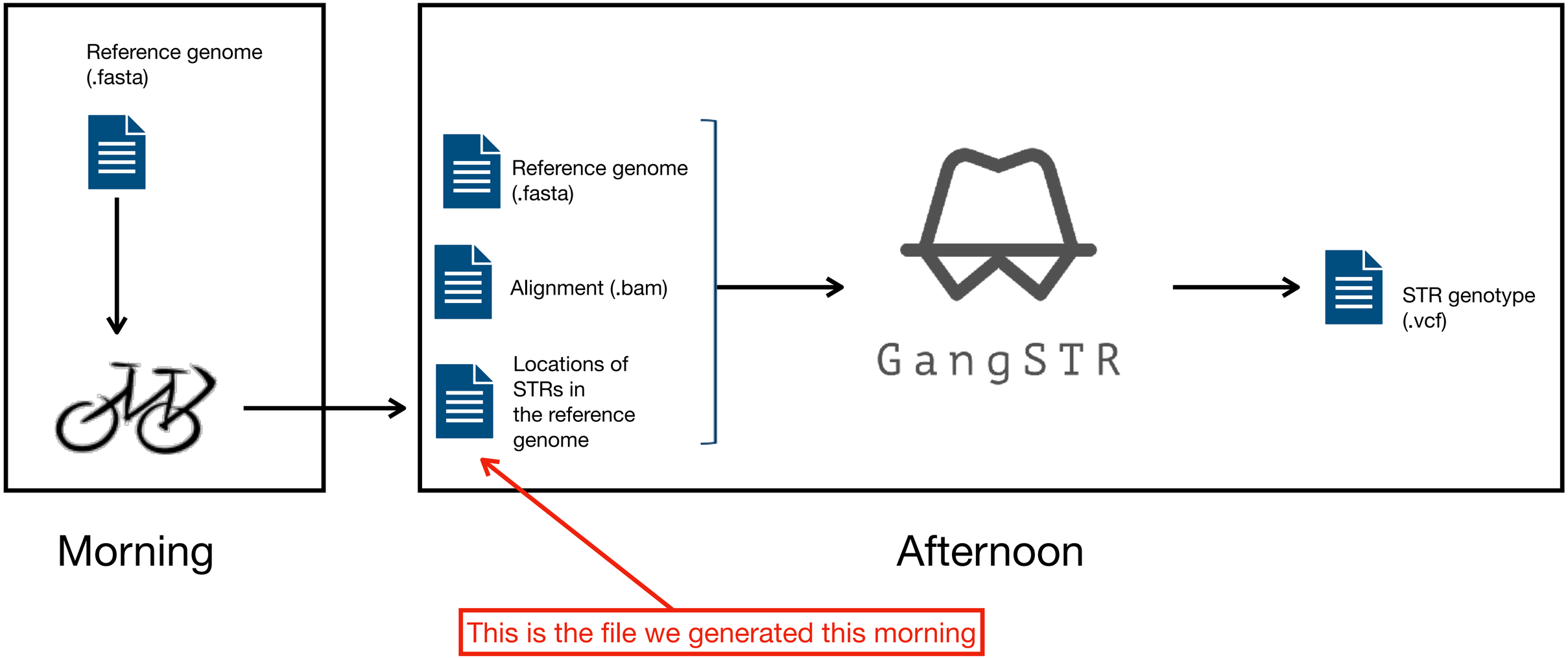
Analysing short tandem repeats in genomic sequences

BIO392

Recap

- Short tandem repeats can mutate rapidly, which can have functional consequences
- Repeat *genotyping*: determining how long repeats are in a given sequencing sample (e.g. using GangSTR)
- Repeat *detection*: determining where repeats are located and how long they are in the reference sequence (e.g. using TRAL)

Workflow for today

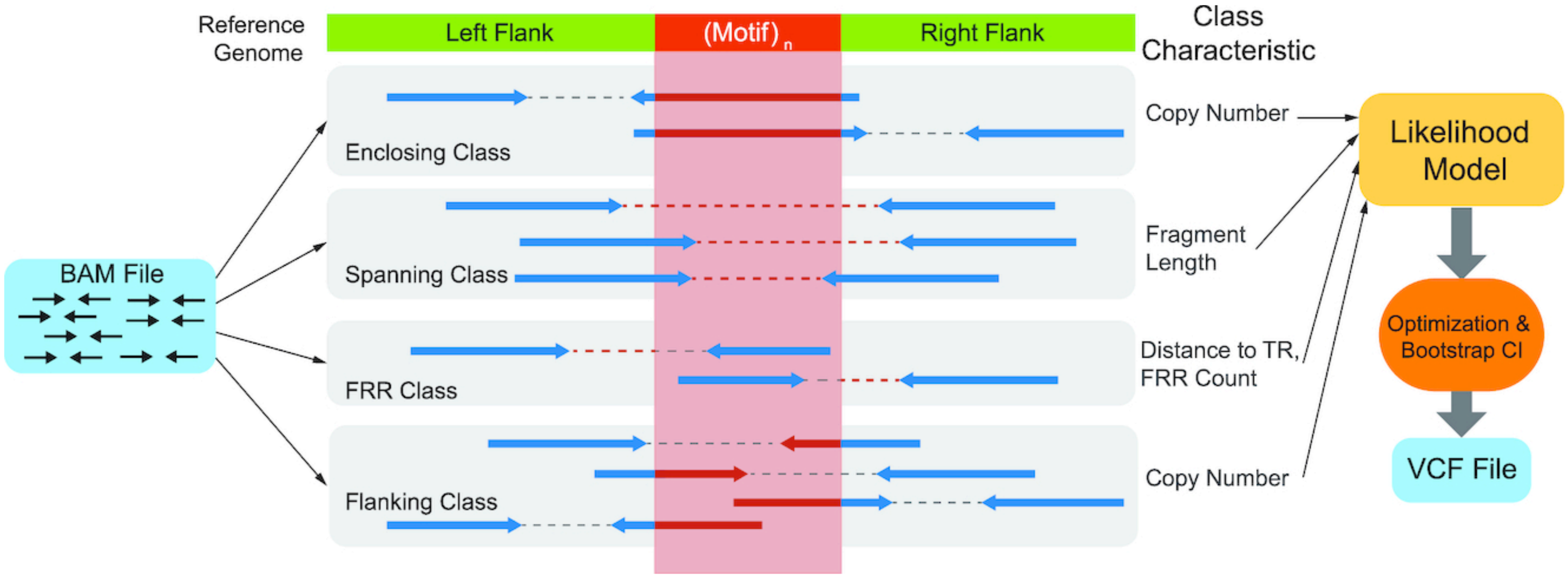


Features of GangSTR compared to other tools

Table 1.
Classes of read pairs and features used by existing tools for genotyping TRs from short reads

Tool	Enclosing	FRR	Spanning	Off-target FRR	Estimates # rpts.	Genome-wide	Estimation limit
LobSTR (18)	X				X	X	< Read length
HipSTR (19)	X				X	X	< Read length
STRetch (26)		X		X	X	X	Only reports expanded TRs
exSTRa (25)		X		X		X	Does not estimate TR length
Tredparse (27)	X	X	X		X		< Fragment Length
ExpansionHunter (28)	X	X		X	X		Poor performance when both alleles long
GangSTR	X	X	X	X	X	X	Not limited by fragment or read length

Using as much information as possible from the alignment



Task 1

GangSTR background reading

Mousavi, Nima, Sharona Shleizer-Burko, Richard Yanicky, and Melissa Gymrek. ‘Profiling the Genome-Wide Landscape of Tandem Repeat Expansions’. *Nucleic Acids Research* 47, no. 15 (5 September 2019): e90–e90. https://doi.org/10.1093/NAR/GKZ501.

- Read the following sections of ‘Profiling the genome-wide landscape of tandem repeat expansions’:
    - Abstract + Introduction
    - Overview of the GangSTR model

- Afterwards, you should be able to answer the following questions:
    - What sets GangSTR apart from other STR genotyping tools?
    - What types of information does GangSTR use for STR genotyping?

Task 2

STR genotyping in the *APC* gene

- Make sure the ‘BIO392_STRs’ conda environment is active
- Start a jupyter server and open ‘scripts/BIO392_GangSTR.ipynb’
- Follow along with the steps in BIO392_GangSTR.ipynb